MISMATCH: Fine-grained Evaluation of Machine-generated Text with Mismatch Error Types

Keerthiram Murugesan [†] **Sarathkrishna Swaminathan**[†] **Soham Dan**[†]

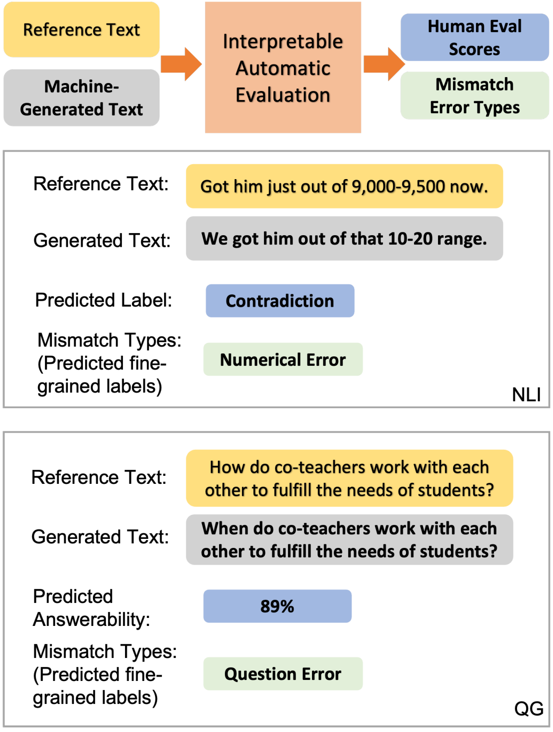
Subhajit Chaudhury[†] **Chulaka Gunasekara**[†] **Maxwell Crouse**[†]

Diwakar Mahajan[†] **Ibrahim Abdelaziz**[†] **Achille Fokoue**[†]

Pavan Kapanipathi[†] **Salim Roukos**[†] **Alexander Gray**[†]

Abstract

With the growing interest in large language models, the need for evaluating the quality of machine text compared to reference (typically human-generated) text has become focal attention. Most recent works focus either on task-specific evaluation metrics or study the properties of machine-generated text captured by the existing metrics. In this work, we propose a new evaluation scheme to model human judgments in 7 NLP tasks, based on the fine-grained mismatches between a pair of texts. Inspired by the recent efforts in several NLP tasks for fine-grained evaluation, we introduce a set of 13 *mismatch error types* such as spatial/geographic errors, entity errors, etc, to guide the model for better prediction of human judgments. We propose a neural framework for evaluating machine texts that uses these mismatch error types as auxiliary tasks and re-purposes the existing single-number evaluation metrics as additional scalar features, in addition to textual features extracted from the machine and reference texts. Our experiments reveal key insights about the existing metrics via the mismatch errors. We show that the mismatch errors between the sentence pairs on the held-out datasets from 7 NLP tasks align well with the human evaluation.



Reference Text: Got him just out of 9,000-9,500 now.

Generated Text: We got him out of that 10-20 range.

Predicted Label: Contradiction

Mismatch Types: (Predicted fine-grained labels) Numerical Error

Reference Text: How do co-teachers work with each other to fulfill the needs of students?

Generated Text: When do co-teachers work with each other to fulfill the needs of students?

Predicted Answerability: 89%

Mismatch Types: (Predicted fine-grained labels) Question Error

Figure 1: Overview of the proposed fine-grained automatic evaluation of machine-generated text with mismatch error types, along with human evaluation scores. Sample examples are taken from Natural Language Inference (NLI) and Question Generation (QG) tasks.

1 Introduction

Large language models have pushed the boundaries for natural language generation (NLG). More and more, the generated machine texts look human-like. The need for evaluation metrics has never been so critical in the recent decade. Typically, there are two ways to evaluate the quality of machine-generated text: automatic evaluation and human evaluation. In automatic evaluation, the quality of the machine-generated text is captured using a single number from a range of values indicating how good the generated text is by a (hand-coded rule-based or neural-based) model. Several NLP tasks still use the metrics from 2 decades ago, for instance, (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) for abstractive summarization, BLEU (Papineni et al., 2002) for machine translation, etc.

It has been noted in several works that automatic evaluation metrics are incapable of capturing the different criteria in measuring the quality of the text and often have a poor correlation with human judgments (Sai et al., 2021; Callison-Burch et al., 2006). The current automatic evaluation metrics lack the ability to measure the quality of a modern machine-generated text. In human evaluation, we evaluate the machine text based on human ratings,

[†] IBM Research. Correspondence to: Keerthiram Murugesan <keerthiram.murugesan@ibm.com>. Code available at https://github.com/IBM/mismatch-eval

where we ask human annotators to judge a given pair of texts. The quality of the machine text is measured using different task-specific human evaluation criteria such as fluency, coherence, correctness, consistency, relevance, adequacy, etc. Human evaluations are often expensive, time-consuming, and subjective (low inter-annotator agreement), especially when broad criteria such as the fluency of the generated text and the interestingness of the model-generated text are used for human judgment.

To address these challenges in automatic and human evaluations, there have been recent efforts in the fine-grained evaluation of generated text in several NLP domains (Callison-Burch et al., 2006; Ethayarajh and Jurafsky, 2020; Sai et al., 2021; See et al., 2019). In this paper, we are interested in utilizing fine-grained evaluation categories to guide the prediction of human judgments. Towards this goal, we introduce a task-agnostic list of 13 *mismatch error types*, such as grammatical errors, spatial/temporal errors, etc, that unifies several related task-specific efforts (Pagnoni et al., 2021; Glockner et al., 2018; Dou et al., 2022). These mismatch error types are comprehensive, interpretable, and useful for predicting human evaluation criteria. For example, an occurrence of grammatical error in a machine-generated text can impact its fluency rating.

Figure 1 gives the overview of the proposed mismatch error types for fine-grained evaluation. We propose a neural framework for evaluation that uses these mismatch error types as auxiliary tasks to model the human judgment and repurposes automated evaluation metrics as additional scalar features, concatenated to textual features extracted from the machine and reference texts via pre-trained LM text embeddings (Devlin et al., 2019). We show that pre-training our proposed model using synthetic data for the mismatch prediction task, and fine-tuning using real data for human evaluation criteria, for different NLP tasks, achieves state-of-the-art performance on the main downstream task of predicting human evaluation metrics. We provide several ablation studies showing the importance of each component of our architecture, and the correlations between the mismatch error types and the automatic and human evaluation metrics. We also show how our architecture is useful in predicting novel evaluation criteria, such as factuality in abstractive summarization.

2 NLG Evaluation

Given a pair of texts: a reference text and a machine-generated one, we are interested in evaluating the quality of the generated text using the reference text. We measure the quality of the generated text by estimating how a human will judge this text based on different evaluation criteria. Such evaluation is common in many ML/NLP tasks, e.g., machine translation, summarization, image captioning, etc. Unlike in other automatic evaluation metrics, we consider fine-grained evaluation cues from 13 mismatch error types, inspired by several related task-specific efforts (Pagnoni et al., 2021; Glockner et al., 2018; Dou et al., 2022) to guide the main task of predicting the human judgments. We propose a neural framework for evaluating machine-generated texts that use these mismatch error type predictions as auxiliary tasks, and automated evaluation metrics as additional scalar features, along with the pair of pre-trained LM text embeddings extracted from reference and generated texts. In this section, we discuss the role of mismatch error types as a good proxy for human judgments (Section 2.1) and the model architecture for the proposed approach (Section 2.2).

2.1 Mismatch Error Types

Recently there has been a growing interest in a set of measurable fine-grained evaluation criteria (Dou et al., 2022; Pagnoni et al., 2021; Glockner et al., 2018). Most of the recent works require human annotation. In this paper, we consider MISMATCH types which identify a specific violation or mismatch between a pair of texts spanning various dimensions of semantic structure: whether the mismatch is within a semantic frame, including predicates, entities, modifiers or across multiple semantic frames, for instance predicate ordering mismatch. These mismatch error types can be used as a proxy to measure the broad evaluation categories: a mismatch in sentence ordering can be a weak signal for the coherence of the generated text, and a change in the object names, gender, and numbers can indicate the correctness of the generated text. Table 1 shows the list of mismatch error types used in this paper. We want to understand the relationships between the mismatch types, the evaluation metrics and the human evaluation criteria by addressing the following three questions:

- *Are mismatch types a good proxy for human evaluation criteria?* We show that the fine-

Error Type	Abbr	**Definition**	**Example Sentence**
Grammatical/Usage Error	GramErr	Faulty or incorrect use of the grammar and syntax.	**ref**: Two paintings are on the wall. **gen**: Two painting is on the wall.
Predicate Error	PredErr	Error in the predicate or its usage with respect to the reference text.	**ref**: John entered the kitchen. **gen**: John found the kitchen.
Entity Error	EntErr	Mismatch in the primary arguments of the predicate.	**ref**: A dog chased a cat. **gen**: A dog chased a rat.
Predicate Ordering Error	PredOrdErr	Error in causal or temporal ordering of the predicates/events.	**ref**: The police arrested the suspect then he was taken to prison. **gen**: The suspect was taken to the prison then the police arrested him.
Hyponyms/Hypernyms Errors	HypErr	Violations in hypernym/hyponym usage.	**ref**: Jim studied mechanical engineering. **gen**: Jim studied architectural science.
Numerical Error	NumErr	Error in numerical, quantifiers or related to numbers (ordinals, cardinals, etc)	**ref**: Martha ate four apples. **gen**: Martha ate six apples.
Spatial/Temporal Error	STErr	Error in spatial or geographic information (location, time, etc).	**ref**: Dave lives in south Chicago. **gen**: Dave lives in south Chile.
Attribute/Modifier Error	AttrErr	Mistakes in additional information concerning the predicates and entities. (not covered by numerical, spatial, geographic)	**ref**: Greg has two small dogs. **gen**: Greg has two big dogs.
Question Error	QuestErr	Error/change in the nature of the question's intention.	**ref**: Did you take the dog to the vet? **gen**: When did you take the dog to the vet?
Negation	NegErr	Negated compared to the reference text.	**ref**: Susan took the gift. **gen**: Susan did not take the gift.
Missing Information	MissInfo	Missing key details from the reference text.	**ref**: Bob drove to the hospital and saw a doctor. **gen**: Bob saw a doctor.
Out of Reference	OutofRef	Contains additional details not present in the reference text.	**ref**: Jack and Jane are friends. **gen**: Jack and Jane are friends. Jack plays football.
Redundant/Repetition	RepErr	Same/similar information repeated more than once.	**ref**: Tom met Sam at the party. **gen**: Tom went to the party. Tom met Sam at the party.

Table 1: Fine-grained evaluation with Mismatch error types between the reference and model-generated texts.

grained evaluation based on the mismatch types can be used to approximate the evaluation criteria used for human ratings.

- *Can we predict a mismatch type between a given pair of texts?* We demonstrate that, in addition to the BERT-based text representations, evaluation metrics computed from the input pair of texts can *reliably* identify these mismatch types (with relatively fewer examples for training).
- *Can we use these evaluation metrics to predict mismatches on an unseen text pair?* We study the predictive power of these evaluation metricsand demonstrate that even though the evaluation metrics do not agree with human evaluation criteria, they can easily identify these mismatch types between pairs of text.

As we later see in Figure 3, our proposed mismatched error types correlate well with both the automatic evaluation metrics as well as human evaluation criteria, demonstrating the relevance of these error types. In the next section, we show how we model the human ratings on 7 NLP tasks: Abstractive Summarization (AS), Image Caption generation (IC), Question Generation (QG), Machine Translation (MT), Dialogue Generation (DG), Data-to-Text generation (D2T) and Natural Language Inference (NLI) using the mismatch types.

2.2 Mismatch Error Types for NLG Evaluation

We now discuss the neural architecture for the proposed NLG evaluation and show how we model the human judgments using the mismatch error types.

A simple solution to model the human judgments is to directly train a neural network to learn a function that maps the input pairs of texts to human ratings on the different evaluation criteria, but the amount of human-annotated samples available for training in many NLP tasks is very limited. In this paper, we consider fine-grained evaluation cues based on mismatch error types to guide the model for predicting human judgments. One of the key advantages of using mismatch error types to approximate human ratings is that we can generate a large amount of synthetic data for these error types. In this paper, we generate $\approx 160K$, synthetic examples for 13 mismatch error types and use publicly available task-specific data with dataset size ranging from a few thousand to hundreds of thousands of examples with human annotation (details in Section 3).

Our approach to model human judgments involves two steps: 1) task-agnostic pre-training step where we use the synthetic examples from mismatch error types to train a shared (base) neural network model for all the 7 NLP tasks and 2) task-specific finetuning step where we finetune the pre-

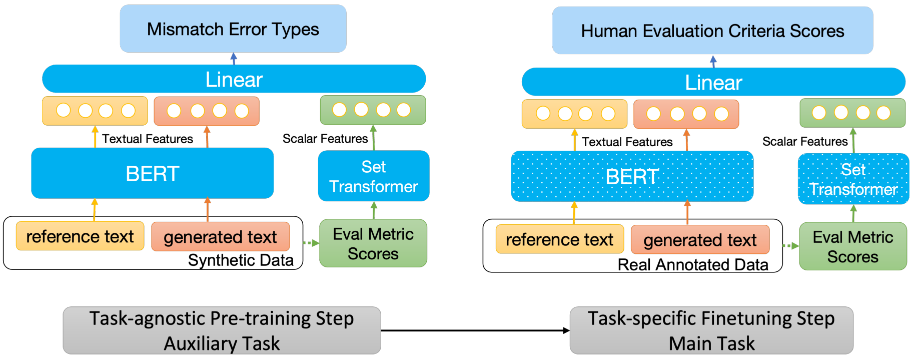


Figure 2: An overview of the mismatch-based evaluation architecture showing task-agnostic pre-training with mismatch error type prediction on synthetic data as the auxiliary task (left) and task-specific finetuning with human evaluation criteria on real annotated data as the main task (right). Dotted arrows indicate that the evaluation metric scores are pre-computed. Dotted blocks indicate that the modules are reused from the pre-training step.

trained model for a specific task to predict the human ratings over different evaluation criteria. The output from the task-specific models approximates the human judgments along with the interpretable mismatches between the given pair of texts.

Figure 2 shows the architecture for the proposed mismatch-based evaluation model with pre-training and finetuning steps. In both these steps, we use pre-trained BERT (Devlin et al., 2018) to extract linguistic features via embeddings for both the reference and generated texts to predict the mismatch types and the human ratings (*textual features*). We generate (2x) 64-dimensional textual features, one for the machine text and the other for the reference text. It is common to generate millions of synthetic examples for pre-training the neural network (Sellam et al., 2020) or to use tens of thousands of human-annotated data (Rei et al., 2020) to make the model robust to unseen texts. On the other hand, automatic evaluation metrics utilize handcrafted logic to compute the score on any pair of texts. In Section 3, we show that evaluation metrics as features can reliably predict these mismatch error types (*scalar features*). We demonstrate that even with a few (synthetic and task-specific) samples, our approach benefits from the handcrafted logic in the evaluation metrics to boost the prediction performance. We choose the evaluation metrics from different NLP tasks to represent different properties of natural language text.

Unlike the textual features, the features from the automatic evaluation metrics are required to be invariant to different permutations. Traditional neural network-based models (including BERT) are very sensitive to the permutations of the input sequence. We use SetTransformer (Lee et al., 2019) to extract permutation-invariant scalar features from the automatic evaluation metrics so that the scalar feature does not change under any permutation of the evaluation metric scores. We scale the evaluation metric scores between 0 and 1 before passing them to SetTransformer. We believe that textual features are extremely useful for the prediction when the reference and/or machine-generated texts are similar to the texts seen during pre-training or finetuning steps whereas scalar features are good for unseen texts. Based on this intuition, we combine the reference and generated texts with the scores computed from the automatic evaluation metrics for prediction. Both the textual and scalar features are concatenated and projected (via linear layer) to either 13 mismatch error types for the pre-training step or human ratings on the task-specific evaluation criteria during the finetuning step.

3 Experimental Results

In this section, we show experimental results validating the proposed model for predicting human judgments.

3.1 Datasets

To train our proposed model based on mismatch error types to predict human judgments, we use synthetic examples for 13 mismatch error types during pre-training and real annotated examples from 7 NLP tasks for task-specific finetuning. We generate the synthetic examples by sampling the reference text from multiple NLP tasks (SQuAD (Rajpurkar et al., 2016), WebNLG (Gardent et al., 2017), MSCOCO (Lin et al., 2014)). Following the previous works (Sai et al., 2021; Glockner et al., 2018), we use template-based perturbations on reference

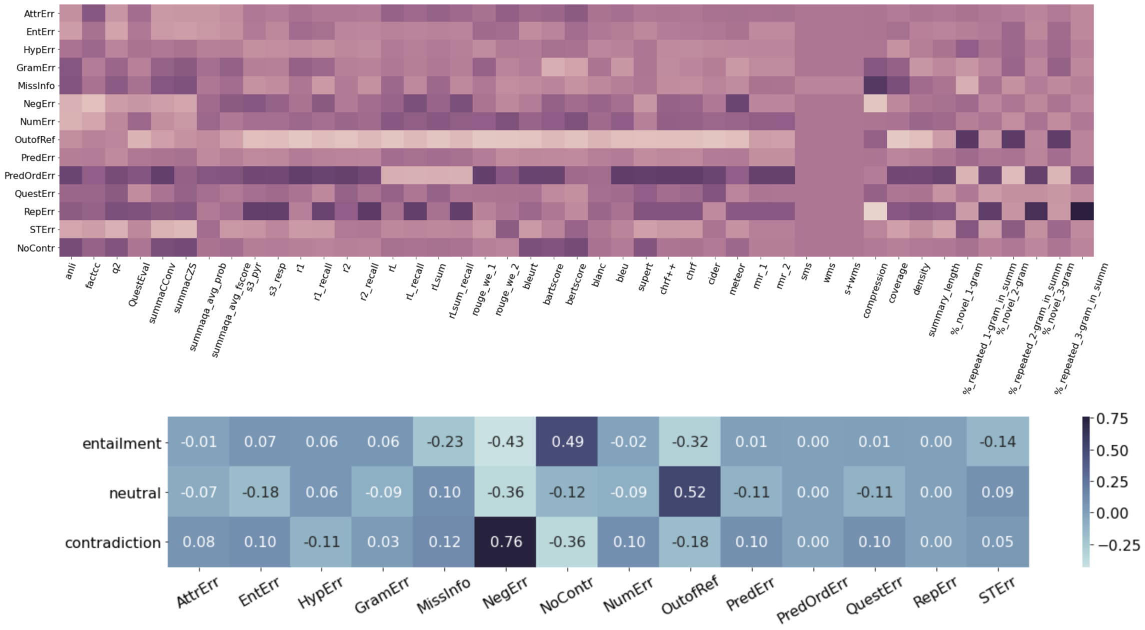


Figure 3: (Top) Correlation between mismatch error types vs automatic evaluation metrics from different NLP tasks. (Bottom) Correlation between mismatch error types vs human evaluation criteria for NLI task.

text to generate the synthetic examples for each mismatch type. E.g., perturbation rules to introduce subject-verb disagreement or dropping stopwords for GramErr, changing names/gender or changing the object order for EntErr, etc. In addition to the 13 error types, we include an additional *No Contradiction* type (NoContr) for machine-generated text that matches the reference text. We believe this additional category helps with the better prediction of mismatch types during pre-training. We generate $\approx 200K$ synthetic examples in total for the pre-training task ($160K$ for training and the rest for validation). We report results on datasets from 7 NLP Tasks with human annotations: AS (Fabbri et al., 2021), IC (Aditya et al., 2015), QG (Nema and Khapra, 2018), MT (Bojar et al., 2017), DG (Mehri and Eskenazi, 2020), D2T (Gardent et al., 2017) and NLI (Williams et al., 2018) for task-specific finetuning step. We show the number of human-annotated examples used for task-specific finetuning for all 7 NLP tasks in Table 2.

3.2 Correlation with Mismatch Error Types

Since most automatic evaluation metrics correlate poorly with human evaluation criteria, we study how well the proposed mismatch error types correlate with the human evaluation criteria and automatic evaluation metrics. Figure 3 shows the correlation plots for the proposed mismatch error types (with NoContr type). The correlations between mismatch types vs automatic evaluation metrics reveal key insights to justify the use of evaluation metrics as scalar features in our model. For instance, OutofRef is negatively correlated but PredOrdErr is positively correlated with most metrics, ngram-based metrics are highly correlated with RepErr, etc. The hardcoded logic-based evaluation metrics are equally correlated with our mismatch types as the neural network-based evaluation metrics. We also show the correlations between the mismatch types and human evaluation criteria for NLI (entailment, neutral, and contradiction). It is interesting to see that NoContr is positively correlated with entailment, NegErr is positively correlated with contradiction. These correlations will help guide the model for better prediction of human judgments on these human evaluation criteria. We also include the correlation plots for other NLP tasks in the supplementary material.

3.3 Model Performance

In this section, we evaluate our model both at the task-agnostic pre-training and task-specific finetuning steps. We use an 80/20 split for both steps. We precompute the automatic evaluation metrics for the pairs of texts in both synthetic and finetuning datasets for faster computation. We use the accuracy to evaluate the performance of the pretrained model on predicting the mismatch types; RMSE (lower is better), Kendall's τ correlation

Tasks	**AS**	**IC**	**QG**	**MT**	**DG**	**DT**	**NLI**
# Samples *(Task-Specific)*	1600 (SummEval)	2007 (Flickr30k)	2726 (AQG)	240,287 (WMT2017-19)	420 (PersonaChat)	5,918 (WebNLG)	9,818 (MNLI)
RMSE	0.18 (0.00)	0.23 (0.00)	0.16 (0.00)	0.19 (0.00)	0.26 (0.00)	0.24 (0.00)	*
Kendall's τ	0.30 (0.01)	0.49 (0.00)	0.52 (0.01)	0.35 (0.00)	0.31 (0.02)	0.39 (0.01)	0.89 (0.00)
Spearman's ρ	0.41 (0.01)	0.62 (0.00)	0.66 (0.01)	0.50 (0.00)	0.40 (0.03)	0.49 (0.01)	0.92 (0.00)

Table 2: Model performance (agreement with human ratings) measured using Root Mean Squared Error (RMSE), Kendall's τ correlation and Spearman's ρ correlation on 7 NLP tasks (averaged over human evaluation criteria). Top row shows the dataset (in parentheses) and number of samples used for each task during finetuning step. * indicates RMSE is not available as the human ratings are defined on three classes: Entailment, Neutral, Contradiction.

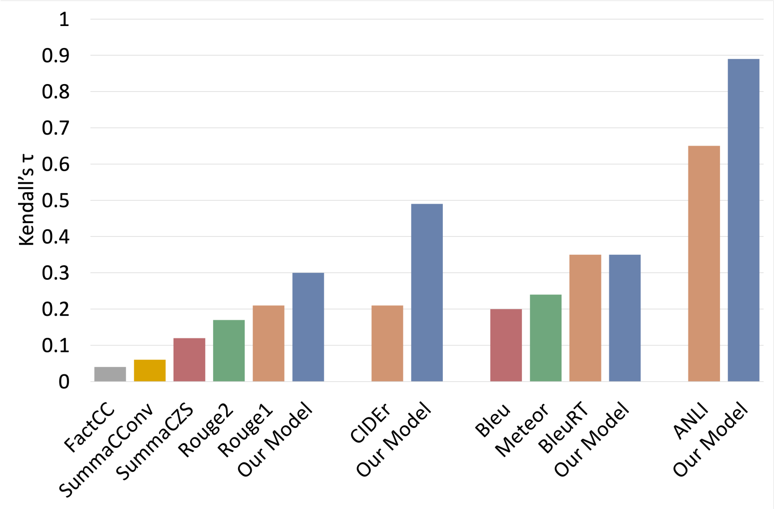


Figure 4: Comparison of the task-specific evaluation metrics with the proposed model. Kendall's τ correlation (agreement with human ratings) is used for the performance comparison. Average over 3 runs with 100 samples randomly taken from the task-specific test set.

(higher is better), and Spearman's ρ correlation (higher is better) between the human ratings and the predicted ratings to evaluate the performance of the task-specific finetuned models. Since we have multiple human evaluation criteria per task (e.g., entailment, neutral, and contradiction in NLI), we report the results by averaging the performance of the finetuned model over the human evaluation criteria from that task. All the experimental results reported in this paper are averaged over 3 random runs.

Table 2 shows the task-specific finetuned model performance on predicting the human rating using both the mismatch error types and scalar features. Our pre-trained model predicts the mismatch types on the held-out synthetic data with 98% accuracy. We finetune the trained model on the task-specific data and achieve relatively lower RMSE scores on most of the tasks. Kendall's τ and Spearman's ρ measure the linear correlation between the human ratings and the model-predicted ratings. We see that in all the NLP tasks, our model predictions align well (≈ 0.50 in correlation) with the human ratings. Since the NLI task involves classification labels (-1 for contradiction, 0 for neutral, and 1 for entailment) instead of human rating scores, we didn't report the RMSE score. We see that the correlation (both τ and ρ) for NLI is high compared to the other tasks. We believe that the NLI task is relatively easier for our proposed model compared to the other task.

Figure 4 compares the proposed model based on the mismatch error types against the task-specific automatic evaluation metrics both hardcoded logic-based and neural network based rules. Kendall's τ correlation between the metrics and the human ratings is used for the performance comparison. We can see that the proposed model outperforms the other metrics significantly in AS, IC and NLI. In addition, we outperform a popular neural network-based evaluation model for machine translation, BLEURT on different language pairs from both WMT2018 and WMT2019 (See supplementary for more details).

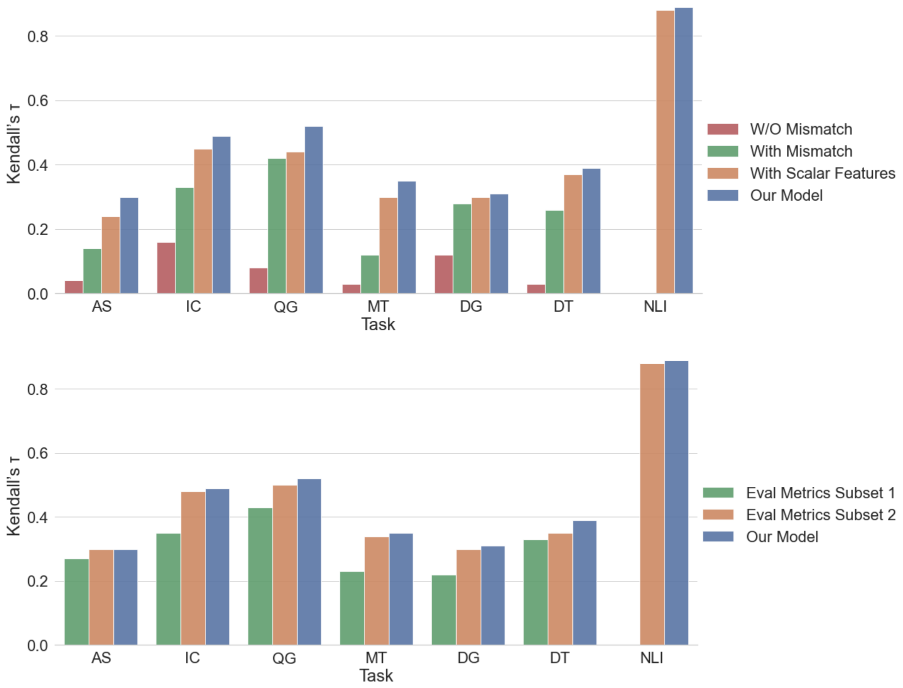


Figure 5: (Top) Agreement on human ratings using Kendall's τ correlation for different settings of the proposed approach on 7 NLP tasks. We start with Textual Features +*Without mismatch* (no pre-training step with mismatch error types) and *Without Scalar Features* (no evaluation metric scores as features during pre-training and finetuning steps) (Bottom) Agreement on human ratings using Kendall's τ correlation with different subset of automatic evaluation metrics used for scalar features. The two subsets are selected based on the overall cost (time and space complexity) to compute the metric scores (See table 11 in the supplementary material).

3.4 Ablation Studies

In this section, we analyze the importance of the evaluation metrics and the mismatch types for predicting task-specific human judgments. First, we compare the proposed model architecture with different settings such as with and without the mismatch error types for pre-training steps and with and without the scalar features extracted from the automatic evaluation scores using SetTransformer. Figure 5 (top) shows Kendall's τ correlation for the different experimental setups. We start with the base model that uses BERT to extract the textual features and predict the human ratings without the pre-training step for predicting mismatch error types and without the scalar features from evaluation metric scores. We write this setup as *Textual Features + Without Mismatch*. Next, the baseline considers the pre-training step with mismatch types but without any scalar features. We write this setup as *Textual Features + With Mismatch*. We consider an additional baseline that uses the scalar features but without the pre-training step for mismatch-type prediction. We call this baseline, *Textual Features+ With Scalar Features*. Finally, we have the proposed model that considers both the prediction step for mismatch error types and scalar features extracted from automatic evaluation metric scores.

We see that in text-only features, the pre-training step with mismatch error type significantly boosts the performance of the task-specific finetuning step, specifically in IC, QG, DG, and DT. In NLI, text-only features didn't perform as well as expected (0.0 Kendall's τ correlation). We observe that using automatic evaluation metric for scalar features significantly boost the performance of the overall model. We believe that evaluation metrics provide valuable properties (both via the hardcoded logic-based metrics such as Rouge, METEOR, etc, and neural network-based evaluation models such as ANLI, FactCC, etc) of the input texts for better performance of the fine-tuned models. The proposed model with both the scalar features and the mismatch error types for pre-training outperforms all the other model setups. Our proposed model gets a little boost from the pre-training with mismatch error types along with the scalar features.

Figure 5 (Bottom) compares the importance of evaluation metric scores as a feature for the model prediction. We know from our previous experiment, evaluation metric score as scalar features provide a significant boost to our proposed model. One of the key issues with using automatic evaluation metrics as features is the cost associated with computing the scores, both space and time complexity. Time complexity measures how long it takes to compute the score for a given pair of text and space complexity measures the storage space occupied by the neural network-based evaluation model. We show the time and space complexity of each metric used in this paper in the supplementary material. To address this concern, we study the importance of cost in our model prediction.

We choose 2 subsets of evaluation metrics with low and high costs. The subset with low-cost metrics includes hardcoded logic-based metrics such as ROUGE, METEOR, etc. The subset with high-cost metrics is mostly neural network-based models such as ANLI, FactCC, SummaC, etc. We observe that metrics with low cost perform comparably to the metrics with high-cost as scalar features. In some tasks such as IC, QG, MT, and DT, the difference is noticeable. In NLI, the difference is significantly higher compared to any other tasks. This reveals that a subset of evaluation metrics can be selected based on the computational constraints to tradeoff between the cost and the model performances.

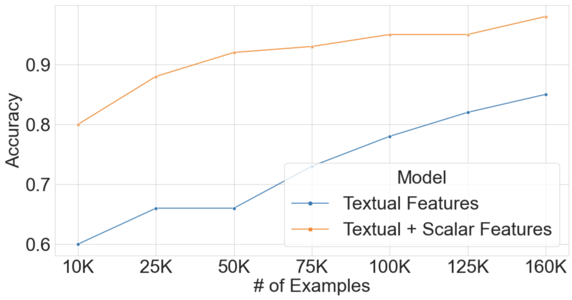


Figure 6: Performance of the pre-trained model with and w/o scalar features on different synthetic sample size. Accuracy of predicting the mismatch error types is used for comparison.

In Figure 6, we study the importance of evaluation metrics as scalar features on sample complexity during the pre-training step. We choose different sample sizes from synthetic data for predicting the mismatch error type ranging from $10K$ to $160K$. We see the model with both the textual and scalar feature achieves better performances with a limited number of samples to train the model. This shows that evaluation metrics as scalar features have likely improved the sample complexity of the proposed model. Finally, in Figure 7, we show some sample text from 3 tasks (IC, QG and DT) showing both the predicted mismatch error type and predicted human evaluation criteria scores.

4 Related Work

Automatic evaluation metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), have been proposed for different tasks as a substitute for human annotations. In NLP, text generation tasks have extensively used these metrics to measure the quality of the machine-generated text. Evaluation metrics are either task-specific (ANLI (Williams et al., 2022), SummaC (Laban et al., 2022), CIDER (Vedantam et al., 2015), SUPERT (Gao et al., 2020)) or task-agnostic (BERTScore (Zhang et al., 2019), BLEURT (Sellam et al., 2020)), and are based on either human handcrafted logic (ROUGE, BLEU, METEOR) or neural framework (BERTScore, BLEURT). For human annotation, several dimensions such as coherence, consistency, and fluency are considered to measure the quality of the generated text, yet most evaluation metrics compute a single score to summarize the evaluation. Further, these single-scored metrics often do not correlate well with the human ratings. To address this problem, several attempts have been proposed to combine multiple evaluation metrics. ROSE (Conroy and Dang, 2008) uses a linear combination of ROUGE variations (ROUGE_1, ROUGE_2, ROUGE_L, ROUGE_Lsum) for the machine translation task and the combined score is better than the individual rouge scores in the evaluation. S^3 (Peyrard et al., 2017) uses the combination of the ROUGE scores and Jenson-Shannon Divergence to predict the human rating. Neural-based approaches such as BLEURT and COMET directly train on the overall human rating for the sample texts. Even though neural-based evaluation metrics seem promising, they often require tens of thousands of training samples to mimic human rating, struggle with new domains/tasks and unseen samples, and still output a single score.

Understanding Evaluation Metrics: Recently, there has been growing interest in understanding what these evaluation metrics measure in terms of fine-grained evaluation criteria. It is done by studying different error categories (mismatches) in the

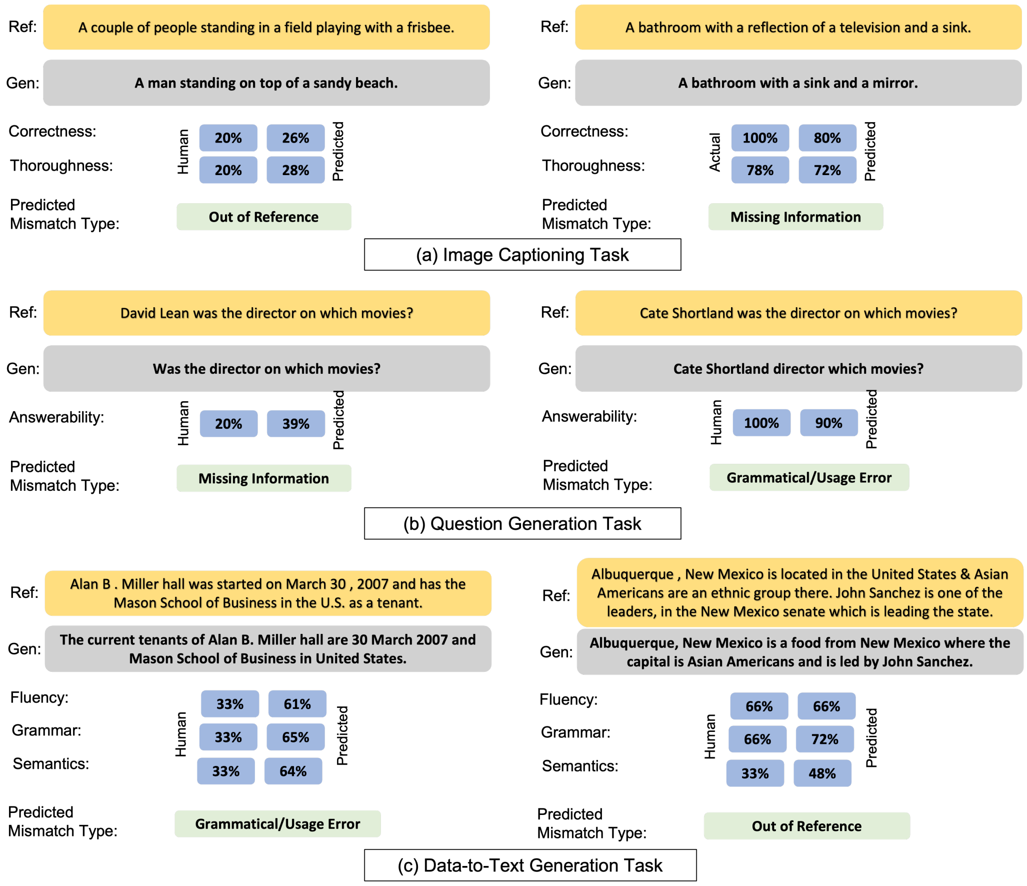

Ref:
A couple of people standing in a field playing with a frisbee.
Gen:
A man standing on top of a sandy beach.
Correctness:
Thoroughness:
Human
20%
26%
20%
28%
Predicted
Predicted Mismatch Type:
Out of Reference
Ref:
A bathroom with a reflection of a television and a sink.
Gen:
A bathroom with a sink and a mirror.
Correctness:
Thoroughness:
Actual
100%
80%
78%
72%
Predicted
Predicted Mismatch Type:
Missing Information
(a) Image Captioning Task
Ref:
David Lean was the director on which movies?
Gen:
Was the director on which movies?
Answerability:
Human
20%
39%
Predicted
Predicted Mismatch Type:
Missing Information
Ref:
Cate Shortland was the director on which movies?
Gen:
Cate Shortland director which movies?
Answerability:
Human
100%
90%
Predicted
Predicted Mismatch Type:
Grammatical/Usage Error
(b) Question Generation Task
Ref:
Alan B . Miller hall was started on March 30 , 2007 and has the Mason School of Business in the U.S. as a tenant.
Gen:
The current tenants of Alan B. Miller hall are 30 March 2007 and Mason School of Business in United States.
Fluency:
Grammar:
Semantics:
Human
33%
61%
33%
65%
33%
64%
Predicted
Predicted Mismatch Type:
Grammatical/Usage Error
Ref:
Albuquerque , New Mexico is located in the United States & Asian Americans are an ethnic group there. John Sanchez is one of the leaders, in the New Mexico senate which is leading the state.
Gen:
Albuquerque, New Mexico is a food from New Mexico where the capital is Asian Americans and is led by John Sanchez.
Fluency:
Grammar:
Semantics:
Human
66%
66%
66%
72%
33%
48%
Predicted
Predicted Mismatch Type:
Out of Reference
(c) Data-to-Text Generation Task

Figure 7: Sample examples taken from 3 tasks (IC, QG and DT) with both predicted mismatch error type and predicted human evaluation scores. Both the human-annotated (human) and our model-estimated (predicted) evaluation criteria scores are reported for comparison.

machine-generated texts but claim none of the existing evaluation metrics can predict all of the mismatches, without providing any solution. Perturbation checklist (Sai et al., 2021) uses template-based perturbation on multiple tasks based on the human evaluation criteria to study mismatches. FRANK (Pagnoni et al., 2021) studies different evaluation metrics on factuality in abstractive summarization using error types. Scarecrow (Dou et al., 2022) explores errors in prompt-based text generation by large language models. BreakingNLI (Glockner et al., 2018) evaluates different metrics on synthetic data created from the external knowledge graph WordNet (Miller, 1995). Tang et al. (2021) studies different types of factuality and hallucinations in the generated text by large language models.

In this work, we unify these task-specific research directions. We propose several evaluation models that combine the best of handcrafted logic on robust evaluation, a neural framework for text representations, and mismatch error types types to measure the quality of the generated text based on the human evaluation criteria.

5 Conclusion

In this paper, we proposed a neural framework for evaluating the quality of the machine-generated text w.r.t the reference text. To achieve this, we defined a set of mismatch error types to approximate the human ratings over a set of evaluation criteria. We showed that in addition to the BERT-based text representation, feature-invariant representations learned from the automatic evaluation metrics improve the prediction of both the mismatch types as well as human ratings with pre-training on only a limited amount of synthetic examples with mismatch error types. We further showed that mismatches between pairs of texts provide an interpretable way to explain human judgments, through a series of ablation studies and correlation analyses. Our proposed mismatch error types is a crucial bridge between automatic evaluation metrics and human evaluation criteria, leading to more interpretable predictions for NLP models.

6 Limitations

One limitation of our work, which is also an avenue for future work, is that it is not fully understood yet why the mismatch error types help much more in some tasks than others. Trying to develop a more task or even instance-specific understanding of the benefits of mismatch error types will be very useful. We also want to try our proposed approach on a wider set of tasks, using different foundational models, and under the distribution shift setting to see if the mismatch error types as auxiliary supervision can improve robustness of natural language processing systems.

7 Ethics Statement

With the ubiquity of natural language processing systems in real-world applications, especially in sensitive domains, it is very important that the machine-generated text is of high quality, as measured by a list of human evaluation criteria such as coherence, consistency, among others. Thus, from a societal perspective, our proposed mismatched error types provides a way to evaluate the quality of machine-generated text with respect to the reference text. From an ecological perspective, our proposed model design only involves synthetic data for pre-training and minimal computation overhead. In addition, from a trustworthiness perspective, MISMATCH provides an interpretable scheme to identify the differences between pairs of text which makes it very suitable for sensitive applications in NLP.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovickỳ, and Tomáš Musil. 2017. Results of the wmt17 neural mt training task. In *Proceedings of the second conference on machine translation*, pages 525–533.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

John Conroy and Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, pages 145–152.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2022. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Alexander Richard Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. Blanc: Learning evaluation metrics for mt. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 740–747.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Preksha Nema and Mitesh M Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959.

Jun Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. 2021. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. Anlizing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733.

A Mismatch Error Types Correlation Plots

Error Type	**Definition**	**Example Sentence**	**Human Evaluation Criteria**
Grammatical/Usage Error	Faulty or incorrect use of the grammar and syntax.	ref: Two paintings are on the wall. gen: Two painting is on the wall.	Fluency
Predicate Error	Error in the predicate or its usage with respect to the reference text.	ref: John entered the kitchen. gen: John found the kitchen.	Answerability, Relevance, Making sense
Entity Error	Mismatch in the primary arguments of the predicate.	ref: A dog chased a cat. gen: A dog chased a rat.	Relevance, Correctness, Thoroughness, Informativeness, Referential clarity
Predicate Ordering Error	Error in causal or temporal ordering of the predicates/events.	ref: The police arrested the suspect then he was taken to prison. gen: The suspect was taken to the prison then the police arrested him.	Flow/Coherence, Answerability, Making sense, repetitions
Hyponyms/ Hypernyms Errors	Violations in hypernym/hyponym usage.	ref: Jim studied mechanical engineering. gen: Jim studied architectural science.	Informativeness
Numerical Error	Error in numerical, quantifiers or related to numbers (ordinals, cardinals, etc)	ref: Martha ate four apples. gen: Martha ate six apples.	Correctness, Thoroughness
Spatial/Temporal Error	Error in spatial or geographic information (location, time, etc).	ref: Dave lives in south Chicago. gen: Dave lives in south Chile.	Correctness, Thoroughness
Attribute/Modifier Error	Mistakes in additional information concerning the predicates and entities. (not covered by numerical, spatial, geographic)	ref: Greg has two small dogs. gen: Greg has two big dogs.	Correctness, Thoroughness
Question Error	Error/change in the nature of the question's intention.	ref: Did you take the dog to the vet? gen: When did you take the dog to the vet?	Answerability
Negation	Negated compared to the reference text.	ref: Susan took the gift. gen: Susan did not take the gift.	Adequacy
Missing Information	Missing key details from the reference text.	ref: Bob drove to the hospital and saw a doctor. gen: Bob saw a doctor.	Adequacy, Thoroughness, Data Coverage
Out of Reference	Contains additional details not present in the reference text.	ref: Jack and Jane are friends. gen: Jack and Jane are friends. Jack plays football.	Adequacy, Making Sense, Listening
Redundant/Repetition	Same/similar information repeated more than once.	ref: Tom met Sam at the party. gen: Tom went to the party. Tom met Sam at the party.	Thoroughness, Avoid Repetition, Data coverage

Table 3: Mismatch Error types between the reference and model-generated texts.

In Table 3, we present our proposed mismatched error types, their definitions, and examples, and the corresponding human evaluation criteria they aim to capture. In Figure 8, we present the correlation plots between the mismatch error types and the human evaluation criteria for 7 popular NLP tasks. We see a significant correlation between several of the mismatch error types with the human evaluation criteria, especially the ones they aim to capture, across the different tasks.

B Comparison with BLEURT on WMT Shared Metric Task

In this section, we compared the proposed neural evaluation framework against BLEURT, a popular neural network-based evaluation metric in Machine Translation. BLEURT is a strong baseline for our proposed approach where they used automatic evaluation metrics such as ROUGE, BLEU, and BERTScore as pre-training signals for the auxiliary task. Unlike our proposed approach for 7 NLP tasks, the BLEURT evaluation metric is primarily used for evaluating generated texts by the machine translation models. BLEURT uses 1.8 million synthetic examples from Wikipedia for pretraining whereas our proposed approach uses $\approx 160K$ synthetic examples from datasets such as SQUAD, WebNLG, MSCOCO, etc. The BLEURT metric relies on the linguistic (text) features extracted from the reference and machine-generated texts, whereas, our proposed approach uses both the text features and the scalar features extracted from the automatic evaluation score.

In Tables 4 and 5, we compare the BLEURT results with our mismatch-based evaluation approach on the 2018 and 2019 WMT Metric Shared Task. We see that on both datasets, we outperform BLEURT on all language pairs in terms of Kendall's Tau correlation. This shows that fine-grained evaluation criteria based on mismatch error types are better auxiliary signals than automatic evaluation metrics.

C Sample Examples with Intrepretable Mismatch Error

In Tables 6, 7 and 8, we show sample examples from the task-specific dataset. One of the advantages of our proposed methods is that in addition to predicting the human rating based on the different human evaluation criteria, it can provide the mismatches occurred between the reference and machine-generated text for further interprtation of the predicted human ratings.

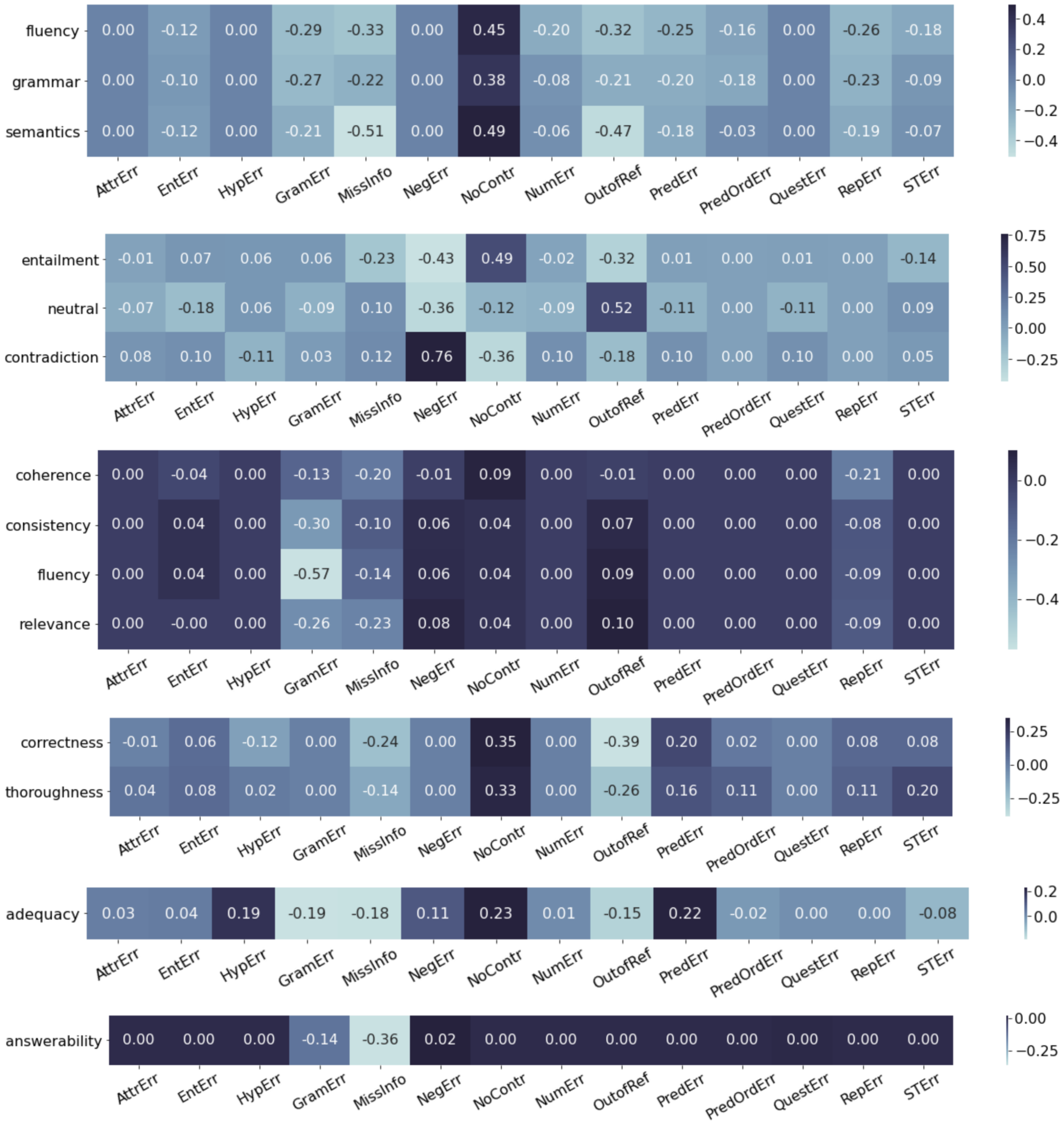


Figure 8: Correlation between mismatch error types vs human evaluation criteria for 7 NLP tasks: Data-To-Text, Natural Language Inference, Abstractive Summarization, Image Captioning, Machine Translation and Question Generation.

D Additional Details

In Table 10, we show the list of all the automatic evaluation metrics used in our proposed model along with the associated NLP task with their references. Table 11 shows the cost associated with computing the evaluation metric scores. It includes both the time complexity (in seconds) and space complexity (in MBs). Time complexity measures how long will the metric takes to compute the evaluation score, whereas space complexity measures how much storage space this metric will consume during the training process. We can see that the hardcoded logic-based metrics such as ROUGE, METEOR, etc are relatively low-cost compared to the neural network-based models such as ANLI, FactCC, etc with the high cost.

Models/ Languages	**cs-en**	**de-en**	**et-en**	**fi-en**	**ru-en**	**tr-en**	**zh-en**	**avg**
BLEURT	35.6	44.2	40.0	32.1	31.9	35.5	29.7	35.6
Our Model	36.0	44.7	40.6	33.3	32.5	36.0	30.4	36.9

Table 4: Agreement with human ratings on the WMT18 Metrics Shared Task. Kendall Tau (τ) is used to evaluate results.

Models/ Languages	**de-en**	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**	**avg**
BLEURT	31.2	31.7	28.3	39.5	35.2	28.3	42.7	33.8
Our Model	31.6	32.3	28.1	40.5	35.4	28.3	45.8	33.9

Table 5: Agreement with human ratings on the WMT19 Metrics Shared Task. Kendall Tau (τ) is used to evaluate results.

Reference Text	**Model Generated Text**	**Correctness**	**Thoroughness**	**Predicted Correctness**	**Predicted Thoroughness**	**Predicted Mismatch**
a mounted police officer riding down a city street past parked cars	a man riding a horse on a city street	1	1	0.82	0.74	Error in Hyponyms or Hypernyms
a couple of people standing in a field playing with a frisbee.	a man standing on top of a sandy beach	0.2	0.2	0.26	0.28	Out of Reference
a bathroom with a reflection of a television and a sink.	a bathroom with a sink and a mirror	1	0.8	0.78	0.72	Missing Information

Table 6: Example sentences from Image Captioning task with predicted human evaluation criteria and mismatch type.

Reference Text	**Model Generated Text**	**Answerability**	**Predicted Answerability**	**Predicted Mismatch**
How do co-teachers work with each other to fulfill the needs of students?	When do co-teachers work with each other to fulfill the needs of students?	1	0.89	Question Error
David Lean was the director on which movies ?	was the director on which movies ?	0.2	0.39	Missing Information
Cate Shortland was the director on which movies ?	Cate Shortland director which movies ?	1	0.90	Grammatical & Usage

Table 7: Example sentences from Question Generation task with predicted human evaluation and mismatch type.

Reference Text	**Model Generated Text**	**Label**	**Prediction**	**Predicted Mismatch**
We got him out of that 10-20 range.	Got him just out of 9,000-9,500 now.	Contradiction	Contradiction	Numerical or Quantifiers
How about when you were in school?	How about when you went to bed?	Contradiction	Contradiction	Entity Error

Table 8: Example sentences from NLI task with predicted human evaluation and mismatch type.

Reference Text	**Model Generated Text**	**Fluency**	**Grammar**	**Semantics**	**Predicted Fluency**	**Predicted Grammar**	**Predicted Semantics**	**Predicted Mismatch**
alan b . miller hall was started on march 30 , 2007 and has the mason school of business in the u . s . as a tenant	the current tenants of alan b. miller hall are 30 march 2007 and mason school of business in united states	0.33	0.33	0.33	0.61	0.65	0.64	Grammatical & Usage
albuquerque , new mexico is located in the united states and asian americans are an ethnic group there . john sanchez , is one of the leaders , in the new mexico senate which is leading the state	albuquerque, new mexico is a food from new mexico where the capital is asian americans and is led by john sanchez.	0.66	0.66	0.33	0.66	0.72	0.48	Out of Reference

Table 9: Example sentences from Data-to-Text task with predicted human evaluation criteria and mismatch type.

Metric	**Task**	**Reference**
ANLI	Natural Language Inference	(Nie et al., 2020)
factcc	Abstractive Summarization	(Kryściński et al., 2020)
Q2	Knowledge-grounded Dialogue	(Honovich et al., 2021)
QuestEval	Abstractive Summarization	(Scialom et al., 2021)
summaC (Conv/CZS)	Abstractive Summarization	(Laban et al., 2022)
summaqa (avg_prob/avg_fscore)	Abstractive Summarization	(Scialom et al., 2019)
S^3 (pyr/resp)	Abstractive Summarization	(Fabbri et al., 2021)
ROUGE	Summarization	(Lin, 2004)
ROUGE-WE	Abstractive Summarization	(Ng and Abrecht, 2015)
BLEURT	Machine Translation	(Sellam et al., 2020)
BARTScore	Text Generation	(Yuan et al., 2021)
Blanc	Machine Translation	(Lita et al., 2005)
Bleu	Machine Translation	(Papineni et al., 2002)
SUPERT	Multi-document summarization	(Gao et al., 2020)
chrf	Machine Translation	(Popović, 2015)
chrf++	Machine Translation	(Popović, 2017)
Cider	Image Description Evaluation	(Vedantam et al., 2015)
Mauve	Text Generation	(Pillutla et al., 2021)
METEOR	Machine Translation	(Banerjee and Lavie, 2005)
RMR (1/2)	Abstractive Summarization	(Zhu et al., 2021)
sms/wms/s+wms	Distance between documents	(Kusner et al., 2015)
coverage/density	Text summarization	(Grusky et al., 2018)

Table 10: Complete list of automatic evaluation metrics used in this paper.

Eval Metrics	**Time Complexity (*sec*)**	**Space Complexity (*MegaBytes*)**
ANLI	426	2163.88
BARTScore	94	1883.47
BERTScore	66	2107.17
Blanc	7048	3081.04
BLEU	47	169.12
BLEURT	869	1949.05
CHRF	80	161.43
CIDER	14	238.25
Datastats (n-gram, etc)	81	399.64
FactCC	165	2197.16
MAUVE	16191	3380.99
METEOR	65	249.09
Q2	18790	6883.66
QuestEval	7996	5493.33
RMR	2012	167.24
ROUGE	1542	181.41
ROUGE_we1	32830	1219.65
ROUGE_we2	34218	1207.79
S^3 *(pyr/resp)*	458	10470.42
SMS	20972	289.70
SummaC_conv	287	2323.56
SummaC_zs	284	2125.90
SummaQA	923	4830.03
SUPERT	532	2237.44

Table 11: The feature costs in terms of the time taken and memory associated with each evaluation metric. Time complexity includes both the data preparation and computation time (in seconds). Space complexity includes average memory taken by the evaluation metrics (in Mb).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.